静态图像中采用混合卷积结构进行人群密度估计 *

范绿源，仝明磊†，李　敏，南　昊

(上海电力学院 电子与信息工程学院，上海 200090)

摘　要：提出了一种混合卷积神经网络用于人群数量的感知计算，在高度密集的场景中可以准确地预测人群密度图。模型仅由两个部分组成：前端为扩张卷积神经网络提取二维特征；后端采用分数步长卷积神经网络降低下采样中的信息损失。为了验证和分析算法性能，模型设计基于当前较为流行的 Shanghai Tech 数据集，使用回归问题的评价指标，即平均绝对误差（MAE）和均方误差（MSE）作为评估算法性能的标准。并且在 Shanghai Tech (MAE=100.8), UCF_CC_50 (MAE=305.3) 与 WorldExpo'10 数据集上进行测试，实验表明模型在密集场景下较以往的方法有效降低了 MAE 和 MSE，提高了密集人群计数的准确率。

关键词：密集场景；扩张卷积；分数步长卷积；密度估计；人群计数

中图分类号：TP391　　doi: 10.19734/j.issn.1001-3695.2018.06.0661

Crowd density estimation using hybrid convolution structure in static images

Fan Lyuyuan, Tong Minglei†, Li Min, Nan Hao

(*School of Electronics & Information Engineering, Shanghai University of Electric Power, Shanghai* 200090, *China*)

Abstract: This paper developed a hybrid convolution neural network for perceptual crowd counting, which could accurately predict density maps in extremely crowded scenes. It consists of merely two components: the front-end is a dilated convolutional neural network to extract two-dimensional features; the back-end deployed a fractionally strided convolution to lower the loss of image information caused by down-sampling. This paper designed the model structure based on the dataset Shanghai Tech, then in an attempt to acknowledge and analyze the performance of the algorithm, , and afterwards made use of the evaluation indicators of the regression problem, the average absolute error (MAE) and the mean-square error (MSE) as the criteria. Additionally, testing the method on Shanghai Tech (MAE=100.8) , UCF_CC_50 (MAE=305.3) and WorldExpo'10 datasets while the experiment results reveal that the proposed model can effectively reduce MAE and MSE when compared with previous methods.

Key words: densely crowded scenes; dilated convolution; fractionally strided convolution; density estimation; crowd counting

0　引言

作为一种人群控制和管理的重要手段，人群密度的精确统计，是当前视频监控领域的一个重要研究方向。某些特定场景人群数量的信息统计，在社会安全、交通流控制方面具有广泛的应用价值。由于人群相互遮挡以及所处环境复杂，现有的方法在实际应用中很难满足要求。卷积神经网络在特征学习中具有显著的性能，可以自动、可靠地获取监测人数或人群密度，报警和预测人群的某些异常行为，而且可以用于人群模拟，人群行为心理学与群体心理学研究。

早期的研究方法[1]往往采用行人检测的方法间接进行，如采用 HOG(histograms of oriented gradients)特征，当人群比较稀疏、人与人之间不存在较大的重叠时，能得到一个比较准确的人数；但当人群变得比较密集时，这种方法得出的结果将不可信。文献[2]提出组合 HOG 特征和颜色直方图特征的检测方法，通过联合两种特征的 SVM（support vector machine）计算结果进行目标判定，消除 HOG 特征检测产生的部分误检。还有一些基于回归的方法，一般通过回归模型如高斯处理回归、线性回归、SVM 回归等求出人群特征与人数之间的函数。该类方法中像素统计特征[3]与人群密度之间的关系较简单，训练后的分类器泛化能力强。但此类方法依赖于提取前景，若前景提取不好则估计效果较差，且密集场景下正确率较低；而基于图像纹理特征[4,5]的人群数量估计方法虽在一定程度上解决了在密集人群中预测效果差的问题，但此方法对稀疏人群估计性能不佳。另外，由于直接在原始图像上提取纹理特征，容易受背景纹理干扰。综上所述传统方法在预测密集人群密度的表现远未达到预期。在卷积神经网络 CNN（convolutional neural network）出色完成各种计算机视觉任务的启发下，许多基于 CNN 的方法得到快速发展，并在人群计数[6,7]方面取得了很大进步，某些基于 CNN 方法如 MCNN（multi-column convolutional neural network）设计多列结构利用不同尺度的感受野提取特征[8]，级联多任务学习 Cascaded-MTL 网络中通过反卷积恢复空间分辨率[9]，极端密集人群图像中利用多源信息[10]回归的人群计数等。但是在感受野限制和图片细节丢失问题方面，算法仍有一定的局限性，虽然在拥挤的环境中通过回归计数是可靠的，但没有

收稿日期：2018-06-30；**修回日期：**2018-08-15　**基金项目：**上海市自然科学基金资助项目（16ZR1413300）

作者简介：范绿源（1994-），女，河南永城人，硕士研究生，主要研究方向为计算机视觉；仝明磊（1976-），男（通信作者），山东郓城人，副教授，博士，主要研究方向为利用人工智能技术的三维视觉重建（Visual Slam 的核心技术）及基于深度对抗网络的路径自动规划方法(tongminglei@gmail.com)；李敏（1994-），女，甘肃陇西人，硕士研究生，主要研究方向为计算机视觉；南昊（1994-），男，浙江乐清人，硕士研究生，主要研究方向为计算机视觉.

对象位置的信息，它们对于低密度人群的预测往往被高估。此类方法的鲁棒性取决于统计数据的稳定性，而在高密度场景中，样本数量往往比较小，不能帮助探索其内在的统计原理。因此，Lempitsky 等人[11]提出了一种在局部通道特征与对应的目标密度图之间进行线性映射的新方法，将图像中存在的空间信息结合起来。最近，Sam 等人[12]提出了一种利用密度等级分类器对特定输入块而选择不同回归函数的可切换式 CNN。这两种解决方案目前达到了最先进的性能，并且两者都使用基于多列的体系结构（MCNN）和密度级分类器[13]。然而以上方法的主要不足体现在：a)多列 CNN 网络较宽，这种扩张的网络结构需要更多的时间进行训练；b)上述的两种解决方案均用到密度级分类器，因为对象数量的大范围变化，在实时拥挤场景分析中，密度水平的粒度难以定义；c)使用分类器意味着需要实现更多的列，这使得设计更加复杂。

考虑到上述存在的问题，本文提出一种在集群场景下编码更广、更深特征的新方法，并生成高品质的密度图。模型通过扩张卷积来增加感受野，同时减少网络参数的数量，并且最终利用分数步长（转置）卷积层来尽可能地恢复细节丢失。

1　密度图生成算法

训练数据中密度图质量决定了人数统计算法的性能。首先将带有标签的人头像转换为人群密度图。如果在像素 x_i 处存在头部，将其表示为 δ 函数 $\delta(x-x_i)$。因此，具有 N 个头标的图像可以表示为一个函数：

$$f(x)=\sum_{i=1}^{N}\delta(x-x_i) \tag{1}$$

为了将其转换为连续密度函数，使用高斯核[11] G_δ 来卷积该函数，使得密度为 $\rho(x)=f(x)*G_\delta(x)$。事实上，每个 x_i 是 3D 场景中地面人群密度的样本，并且由于透视失真，像素与不同样本相关的 x_i 对应于场景中不同大小的区域。因此，为了准确估计人群密度 ρ，需要考虑由地平面与图像平面之间的单应性引起的失真。假设在每个头部周围，人群相对均匀分布，那么头部和其最近的 k 个头部（在图像中）之间的平均距离给出几何失真的合理估计（由透视效应引起）。因此，应该根据图像内每个人的头部大小来确定扩散参数 σ。

很多情况下由于遮挡而难以精确地获得头部的大小，找到头部大小与密度图之间的基本关系也非常困难。通常头部大小与拥挤场景中两个相邻头部中心之间的距离有关，所以根据其与最近邻的平均距离自适应地确定每个人的参数。对于给定图像中的每个头 x_i，将到它的 k 个最近邻的距离表示为 $\{d_{1i},d_{2i},\dots,d_{ki}\}$。平均距离为

$$\bar{d}_i=1/k\sum_{j=1}^{k}d_{ji} \tag{2}$$

因此，与 x_i 相关的像素对应于 x_i 上的区域。为了估计像素 x_i 周围的人群密度，需要将 $\delta(x-x_i)$ 与具有与 d_i 成比例的方差 σ_i 的高斯核进行卷积，更确切地说，密度 ρ 应该是

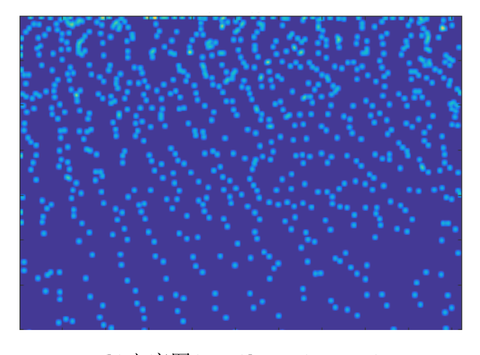
$$\rho(x)=\sum_{i=1}^{N}\delta(x-x_i)*G_{\sigma_i}(x),\quad with\ \sigma_i=\beta\bar{d}_i \tag{3}$$

对于某些参数 β，将标签 f 与密度内核进行卷积，该密度内核适应于每个数据点周围的局部几何分布，称为几何自适应内核。文献[8]中的经验值 $\beta=0.3$ 给出了最好的结果。模型生成的密度估计图需要与数据集的真实密度图作比较，产生的误差损失反向传播给网络，使训练向损失减小的梯度方向进行，标签密度图的准确率很大程度上影响模型的可行性，效果如图 1 所示。其中图 1（a）为原始图片；（b）为相应的原始密度图，密度图左上角为标签人数。



(a)原始图片

(a) Original image



(b)密度图(sum(density)=817.0)

(b) Density map(sum(density)=817.0)

图 1　图片到密度图

Fig. 1　Image to density map

2　网络结构

首先网络整体配置为，输入一张图像，输出对应的人群密度图（如每平方米有多少人）；然后通过积分获得人数。设计的基本思想是部署双列扩张卷积，用于捕获具有较大感受野的高级别特征，并且生成高质量的密度图而不是粗略地扩展网络复杂度。在本章中首先介绍了提出的体系结构，然后给出了相应的训练方法。

2.1　多元卷积结构配置

在文献 [14] 中，其模型的前端输出图片大小是原始输入的 1/8；继续堆叠更多的卷积层和池化层，输出大小将进一步缩小，并且很难生成高质量的密度图，所以模型后端采用扩张卷积层，用于提取更深的显著性信息以及提高输出分辨率。受该文献的启发，本文模型采用扩张卷积作为网络的前端，因其可以通过增大感受野获取更丰富的特征，文献 [14] 是在分辨率已经降到很低的情况下再使用扩张卷积捕捉更多特征，本文则是先利用扩张卷积获取更多的图像信息。再利用转置卷积将图像尺寸增大为原来的 2 倍。由于设计的模型层数少、结构简单，经过前端网络（含两个池化层）后仅为原始输入的 1/4，再经过转置卷积恢复为原图大小的 1/2。

先将图像样本下采样降低分辨率，然后再用上采样还原回来。该过程使下采样的样本分辨率降低，再上采样后分辨率虽不会得到提升，但是这样可以将小目标分辨率低和面积小的问题还原。因此应用分数步长卷积层作为后端对前端输出层进行上采样，可以从一定程度上补充下采样造成的图片细节损失。模型前端基于 MCNN[8]的分支结构，并在卷积核中加入扩张率参数以此来加大感受野。为了减少网络参数，选取效果最好的模型，将四种类型的双列卷积组合作为实验对象，测试将在第 3 章给出详细讨论，最终选择一个相对较

好的模型，考虑到模型的稳定性，即选择 MSE 最低的模型（MAE 非最小值）。网络的整体结构如图 2 所示。

为了简化网络，除了卷积核的大小和数量，对所有卷积列使用相同的结构（conv - pool - conv - pool）。最大池化作用于每个 2×2 区域，并且采用整流线性单元（parametric rectified linear unit，PReLU）作为激活函数。

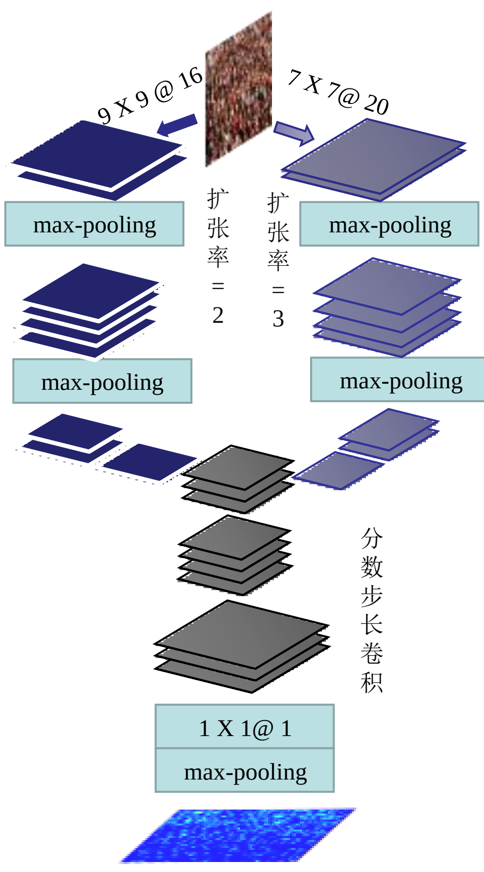


图 2　多元卷积网络结构

Fig. 2　Hybrid convolution network structure

由于透视失真，图像通常包含尺寸相异的头部，具有相同尺寸感受野的卷积核不可能捕获不同尺度的人群密度特征。因此使用具有不同大小的局部感受野的卷积核来学习从原始图像到密度图的映射，即具有较大感受野的卷积核在较大头部对应的密度图上进行建模更有效；同时为了减少计算复杂度（要被优化的参数的数目），使用较少数量的卷积核用于具有较大卷积核的卷积层。例如，网络前端扩张率为 2 的卷积列，9×9 大小的卷积核取 16 个，而 7×7 的取 32 个。后端结合前端卷积层的输出，通过转置卷积层还原图片大小，并采用大小为 1×1 的卷积核进行卷积用于生成高质量密度图。网络配置的详细介绍见表 1。所有的卷积层都使用填充（padding）来保持以前的大小不变。卷积层的参数表示为“conv(kernel size) @ (number of filters)”，最大池化层在步长 2 的 2×2 像素窗口上进行，转置（分数步长）卷积层表示为“ConvTransposed(kernel size) @ (number of filters)”，PReLU 被用作非线性激活层。

然后使用欧氏距离测量真实值与预测密度之间的差异。损失函数定义如下：

$$L(\theta)=\frac{1}{2N}\sum_{i=1}^{N}\left\|Y(X_i;\theta)-Y_i^{GT}\right\|_2^2 \tag{4}$$

其中：θ 是网络中一组可学习参数；N 是训练图像的数量；X_i 是输入图像；Y_i 是图像的真实密度；$Y(X_i;\theta)$ 代表由模型预测密度，其随样本与参数而变化；L 是预测密度与真实密度之间的损失。

表 1　网络参数配置

Table 1　Parameters of network configuration

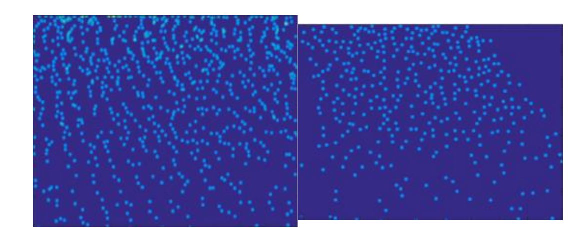
Front-end(double-column)		Back-end(No Dilation)
Dilation rate =2	Dilation rate =3	Conv3x3 @ 24
Conv9x9 @ 16	Conv7x7 @ 20	Conv3x3 @ 32
Max-pooling	Max-pooling	ConvTranspose4x4@16
Conv7x7 @ 32	Conv5x5 @ 40	PReLU
Max-pooling	Max-pooling	Max-pooling
Conv7x7 @ 16	Conv5x5 @ 20	Conv1x1 @ 1
Conv7x7 @ 8	Conv5x5 @ 10	Max-polling

模型最终生成密度估计图的效果如图 3（c）所示。每列图由上到下分别是原始图片、对应人群密度图、预测人群密度图（输入图片来自 Shanghai tech 数据集[8]）。图 3（a）左的原始图像尺寸为 1024x768，相对应的标计人数（gt_count）817，估计人数（et_count）834；图 3（a）右的原始图像尺寸为 1024x687，对应的标记人数 361，估计人数 355。



(a)原始图片

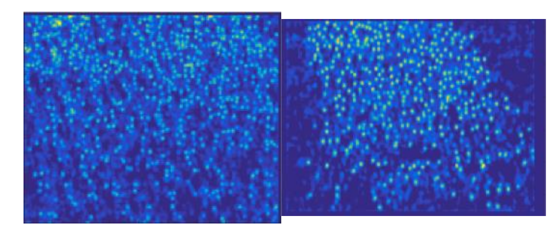
(a) Original image



gt_count:817　　gt_count:361

(b)对应人群密度图

(b) Corresponding crowd density maps



et_count:834　　et_count:355

(c)预测人群密度图

(c) Estimated crowd density maps

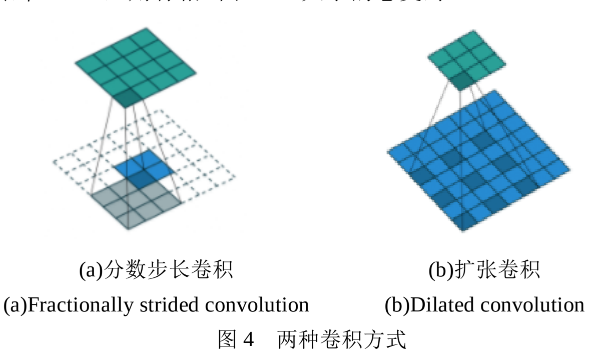
图 3　原始密度图与生成密度图对比

Fig. 3　Comparison of original and generated density maps

本文设计了一个易于训练的基于深度卷积神经网络的密度图管理器。模型使用纯卷积层作为核心，支持灵活分辨率的图像输入。模型利用空洞（扩张）卷积层作为前端以增大感受野而分数步长卷积层作为后端，以恢复其空间分辨率。利用这种简单的结构，降低了网络参数的数量，方便了模型的训练。此外，在 Shanghai tech 数据集 part_A 和 part_B 上，计数结果优于之前人群计数解决方案中的 MAE。

2.2 扩张卷积和转置卷积

模型输入为任意大小图片，输出为人群密度图。网络结构由两个主要部分组成：第一部分学习大尺度特征，第二部分恢复图像大小。网络层越高，单位像素中原始图像所包含的信息越多，也就是感受野越大，通过池化合并完成，代价是原始图像中的信息的减少和丢失。由于池化的存在，后层中的特征映射的大小会越来越小，采用分数步长卷积（转置卷积）将特征图的尺寸变大，一定程度上补充池化层造成的图片细节损失，如图 4（a）所示。扩张卷积是一种卷积的思想，在不增加参数数量或计算量的情况下扩大感受野。在扩张卷积中，具有 k×k 的小尺寸核被扩大为 k+(k-1)(r-1)，扩张率为 r，它允许灵活聚合多尺度信息并保持相同的分辨率，如图 4（b）所示。如果正常卷积核（扩张率＝1）尺寸是 3×3，则其感受野也是 3x3 大小，卷积核尺寸为 3x3 扩张卷积（扩张率＝2），则有相当于 5×5 大小的感受野。



(a)分数步长卷积　(b)扩张卷积

(a)Fractionally strided convolution　(b)Dilated convolution

图 4　两种卷积方式

Fig. 4　Two types of convolution

通过对以下定义的每个人的位置为中心的 2D 高斯内核进行求和来计算与第 i 个训练图像块对应的真实密度图：

$$D_i(x)=\sum_{x_g\in P}G(x-x_g,\delta) \tag{5}$$

其中：σ 是二维高斯核的尺度参数；P 是人群位置的所有点的集合。模型使用 Torch 框架[15]在 NVIDIA TITAN-X GPU 上进行训练和评估。其中 Adam 学习率为 0.000 01，momentum（动量）为 0.9。

3　实验结果

本文分别在三个公开可用的数据集 Shanghai tech、UCF_CC_50[16]以及 WorldExpo'10[6]进行实验。评价指标使用了许多现有人群计数方法所使用的标准，平均绝对误差（mean absolute error，MAE）与均方误差（mean-square error，MSE）度量。标准定义如下：

$$MAE=\frac{1}{N}\sum_{1}^{N}(|y_i-\hat{y}_i|) \tag{6}$$

$$MSE=\sqrt{\frac{1}{N}\sum_{1}^{N}(|y_i-\hat{y}_i|)^2} \tag{7}$$

其中：N 是测试样本数；y_i 是数据集图片中实际标记人数；$\hat{y}_i$ 相应的估计人数。粗略地说，MAE 表示估计的准确性，MSE 表示估计的鲁棒性。

本文将具用不同扩张率的四种组合进行比较。Type1 为第 1 列（扩张率= 2）与第 2 列（扩张率= 3）的组合；Type2 为第 2 列与第 3 列（扩张率= 4）的组合；Type3 组合了第 1 列与第 3 列。Type4 组合了所有列。四种组合方式结构如图 5 所示。

实验将四种组合方式的模型进行分别训练，最终测试结果如表 2 所示。其中 Type3 在 Part_A 部分表现出最好的预测能力，但模型稳定性略低于 Type1。分析原因如下：Type3 与 Type1 配置区别仅在于第 2 列（7x7，5x5）与第 3 列（5x5，3x3）的卷积核尺寸，Type3 中卷积核尺寸小，对于提取较小目标的特征效果好，适于某些人群极端密集，人头部较小的图片，因此 MAE 相对低，但 Type1 更具广泛适用性，对整个数据集的预测估计均表现良好。考虑到模型鲁棒性的重要意义，最终选择较稳定的模型。测试结果如表 2 所示。

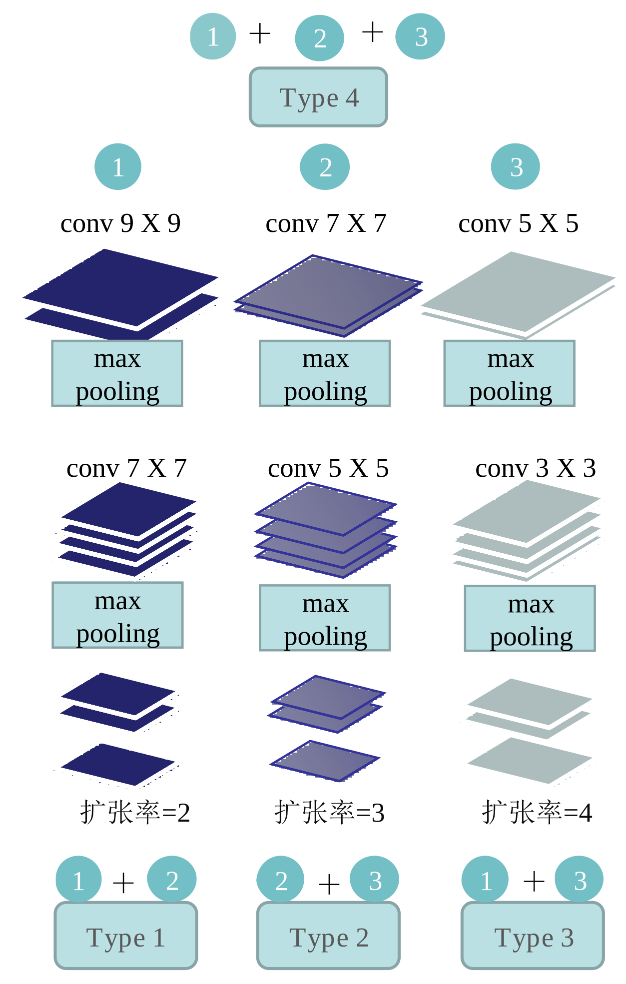


图 5　四种结构组合

Fig. 5　Four types of configuration

表 2　不同组合在 Shanghai tech 数据集上实验结果对比

Table 2　Comparison of experiments results on Shanghai tech dataset

Type	Part_A		Part_B	
	MAE	MSE	MAE	MSE
Tpye1	100.87	152.31	21.55	38.07
Tpye2	103.01	161.98	24.82	45.81
Tpye3	99.66	155.0	28.35	48.78
Tpye4	101.19	160.53	24.15	45.76

3.1　Shanghai tech 数据集

该数据集包含 1 198 图片，共 330 165 人。数据集由两部分组成：Part_A 部分中有 482 幅图从互联网上随机获取，Part_B 部分中 716 幅图从上海大都市地区的繁忙街道拍摄。人群密度在两个子集之间变化很大，使得在人数估计上比大多数现有数据集更具挑战性。

A 和 B 部分均包括训练和测试集：A 部分的 300 张图片用于训练，剩下的 182 张图片用于测试；B 部分的 400 张图片用于训练，316 张用于测试。其中训练集是由每张图像选取不同位置的九个图像块共同组成，尺寸为原始图像的 1/4 大小。前四个图像块包含四个不重叠的图像，而其他五个图像块从输入图像中随机裁剪。模型测试结果如表 3 所示。

表 3　Shanghai tech 数据集密度估计误差对比.

Table 3　Comparison of estimation errors on Shanghai tech dataset.

Method	Part_A		Part_B	
	MAE	MSE	MAE	MSE
Zhang et al[6]	181.8	277.7	32.0	49.8
Marsden et al. [18]	126.5	173.5	23.8	33.1
MCNN[8]	110.2	173.2	26.4	41.3
Switching-CNN[12]	90.4	135.0	21.6	33.4
Cascaded-MTL[9]	101.3	152.4	20.0	31.1
FMFCNN[19]	105.4	168.5	21.7	32.4
本文	100.8	152.3	21.5	33.4

3.2　UCF_CC_50 数据集

UCF_CC_50 数据集包括 50 个具有不同视角和分辨率的图像。每幅图像的标记人数从 94 到 4 543，平均人数为 1 280。按照文献[18]中的标准设置执行 5 折交叉验证。测试结果如表 4 所示。

表 4　UCF_CC_50 数据集密度估计误差对比

Table 4　Compariosn of estimation errors on UCF_CC_50 dataset

Method	MAE	MSE
Zhang et al.	467.0	498.5
MCNN	377.6	509.1
Marsden et al.	338.6	424.5
Cascaded-MTL	322.8	397.9
Switching-CNN	318.1	439.2
本文	305.3	429.4

3.3　WorldExpo'10 数据集

WorldExpo'10 人群统计数据集首次由 Zhang 等人引入。该数据集包含 1 132 个带标签的视频序列，视频均来自 2010 年上海世博会。作者提供了总计 199 923 名行人标签，其中训练集有 3 380 帧，共 103 个场景，每个场景有相应的透视图数据；测试集包括五个不同的视频序列，每个视频序列包含 120 个标注的帧，并为测试场景提供了五个不同的感兴趣区域（ROI）。与前两个数据集不同的是，该数据集提供透视图集，且人群密度分布核函数包含两个项，头部为标准化的 2D 高斯核表示，其余身体部分为二元正态分布函数。按照 文献[6] 的工作根据透视图与 $\sigma = 0.2 * m(x)$ 的关系来生成密度图，$m(x)$ 表示图片中代表该位置一平方米的像素数量。选取其中两个场景进行测试，结果见表 5。

表 5　WorldExpo'10 数据集密度估计误差对比

Table 5　Compariosn of estimation errors on worldexpo'10 dataset

Method	Sce1	Sce5
Zhang et al.	9.8	3.7
Shang et al.[17]	7.8	5.8
MCNN	3.4	8.1
Switching-CNN	4.4	5.9
CP-CNN[20]	2.9	5.8
本文	3.3	4.2

3.4　实验总结

如表 6 所示，将三个基准数据库的参数进行比较总结，Num 是图片的数量，Max 是最大人数，Min 是最少人数，Ave 是平均人数，Total 是已标记人员的总数。

在第 2 章中已经给出 Shanghai tech 数据集高中密度人群的图片测试结果，这里进一步给出其他两个数据集的检测效果。从 UCF_CC_50 数据集测试结果中选取高中低密度人群图片及其检测效果，如图 6 所示，三列图片从左至右依次为原图片、真实密度图及预测密度图，图片最下方对应给出标计人数（gt_count）、估计人数（et_count）。

表 6　基准数据库参数

Table 6　Parameters of benchmark

数据集		Num	Max	Min	Ave	Total
UCF_CC_50		50	4543	94	1280	63974
WorldExpo'10		3980	253	1	50	199923
Shanghai tech	Part_A	482	3139	33	501	241677
FMFCNN	Part_B	716	578	9	124	88488

从 WorldExpo'10 数据集测试结果中选取高中低密度人群图片及其检测效果如图 7 所示。由于该数据集生成密度图中人不只是用一点标记头部，还包括躯体部分。与其他数据集生成密度图略有不同，但训练与测试参数完全相同。图 7 中三列图片从左至右依次为原图片（来源于 WorldExpo'10 的 sce1 与 sce5）、真实密度图及预测密度图，图片最下方对应给出标计人数（gt_count）、估计人数（et_count）。

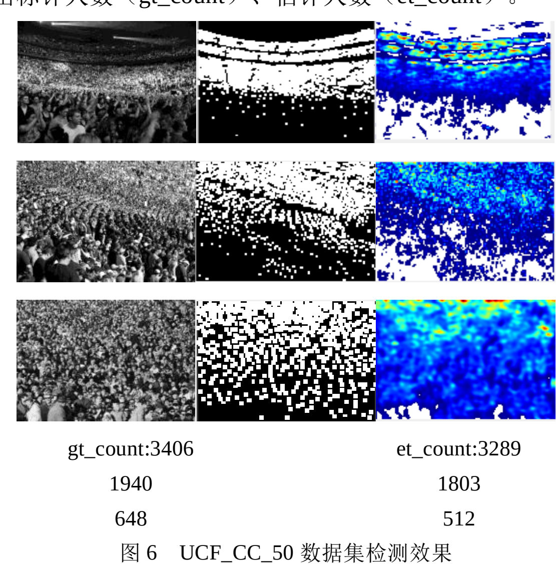


图 6　UCF_CC_50 数据集检测效果

Fig. 6　Estimation performance on UCF_CC_50 dataset

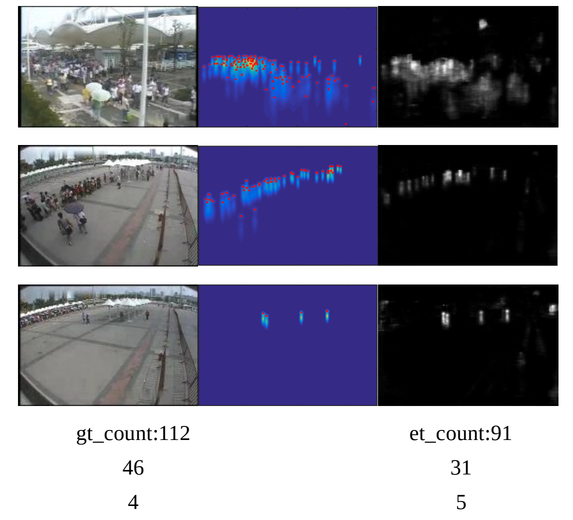


图 7　WorldExpo'10 数据集检测效果

Fig. 7　Estimation performance on worldexpo'10 dataset

4　结束语

本文提出一个扩张卷积与分数步长卷积相结合的多元卷积神经网络。扩张卷积致力于扩大感受野将丰富的特性纳入网络，使其能够更好地学习全局特征，从而统计数据集中的大计数变化。此外，采用分数步长卷积层作为后端，以恢复

在前期阶段经过最大池化层造成的图片细节损失。通过在多个数据集上的测试，本文模型在密集人群的密度估计上具有较好的密度估计性能，模型结构复杂程度适中且泛化能力强，具有普遍适用性。

参考文献：

[1] 张君军，石志广，李吉成. 人数统计与人群密度估计技术研究现状与趋势 [J]. 计算机工程与科学, 2018, 40 (2): 282-291. (Zhang Junjun, Shi Zhiguang, Li Jicheng. Current researches and future perspectives of crowd counting and crowd density estimation technology [J]. Computer engineering and science, 2018, 40 (2): 282-291.)

[2] 高静伟. 通济桥高密度人群计数方法研究与实现 [D]. 广州：中山大学, 2013. (Gao Jingwei. Research and implementation of people counting in high crowd scenes of Tongji Bridge [D]. Guangzhou: Sun Yat-sen University, 2013.)

[3] 王强，孙红. 基于像素统计和纹理特征的人群密度估计 [J]. 电子科技, 2015, 28 (7): 129-132. (Wang Qiang, Sun Hong. Crowd density estimation based on pixel and texture [J]. Electronic Science and Technology, 2015, 28 (7): 129-132.)

[4] Marana A, Costa L D, Lotufo R, *et al.* On the efficacy of texture analysis for crowd monitoring [C]// Proc of IEEE International Symposium on Computer Graphics, Image Processing, and Vision. 1998: 354-361.

[5] Rahmalan H, Nixon M S, Carter J N. On crowd density estimation for surveillance [C]// Proc of Institution of Engineering and Technology Conference on Crime and Security. 2007: 540-545.

[6] Zhang Cong, Li Hongsheng, Wang X , *et al*. Cross-scene crowd counting via deep convolutional neural networks [C]// Proc of Computer Vision and Pattern Recognition. 2015: 833-841.

[7] Boominathan L, Kruthiventi S S S, Babu R V. CrowdNet: a deep convolutional network for dense crowd counting [C]// Proc of ACM on Multimedia Conference. 2016: 640-644.

[8] Zhang Yingying, Zhou Desen, Chen Siqi, *et al*. Single-image crowd counting via multi-column convolutional neural network [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]:IEEE Computer Society, 2016: 589-597.

[9] Sindagi V A, Patel V M. CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting [C]// Proc of IEEE International Conference on Advanced Video and Signal Based Surveillance. 2017: 1-6.

[10] Idrees H, Saleemi I, Seibert C, *et al.* Multi-source multi-scale counting in extremely dense crowd images [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]:IEEE Computer Society, 2013: 2547-2554.

[11] Lempitsky V S, Zisserman A. Learning to count objects in images [C]// Proc of International Conference on Neural Information Processing Systems. [S.l.]:Curran Associates Inc, 2010: 1324-1332.

[12] Sam D B, Surya S, Babu R V. Switching convolutional neural network for crowd counting [C]// Proc of IEEE International Conference on Computer Vision and Pattern Recognition. [S.l.]:IEEE Computer Society, 2017: 5744-5752.

[13] 薛翠红，于洋，张朝，等. 融合 LBP 与 GLCM 的人群密度分类算法 [J]. 电视技术, 2015, 39 (24): 7-10. (Xue Cuihong, Yu Yang, Zhang Zhao, *et al*. Fusing LBP and GLCM for crowd density classification algorithm [J]. Tv Engineering, 2015, 39 (24): 7-10.)

[14] Li Yuhong, Zhang Xiaofan, Chen Deming. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes [C]// Proc of IEEE International Conference on Computer Vision and Pattern Recognition. 2018.

[15] Collobert R, Kavukcuoglu K, Farabet C. Torch7: a matlab-like environment for machine learning [C]// Proc of NIPS Workshop. 2011.

[16] Idrees H, Saleemi I, Seibert C, *et al.* Multi-source multi-scale counting in extremely dense crowd images [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]:IEEE Computer Society, 2013: 2547-2554.

[17] Shang Chong, Ai Haizhou, Bai Bo. End-to-end crowd counting via joint learning local and global count [C]// Proc of IEEE International Conference on Image Processing. 2016: 1215-1219.

[18] Marsden M, McGuiness K, Little S, *et al.* Fully convolutional crowd counting on highly congested scenes[J]. arXiv preprint arXiv: 1612. 00220, 2016.

[19] 唐斯琪，陶蔚，张梁梁，等. 一种多列特征图融合的深度人群计数算法 [J]. 郑州大学学报 ：理学版 , 2018, 50 (2): 69-74. (Tang Siqi, Tao Wei, Zhang Liangliang, *et al*. A deep crowd counting algorithm based on multi-column feature map fusion [J]. Journal of Zhengzhou University :Natural Science Edition , 2018, 50 (2): 69-74.)

[20] Sindagi V A , Patel V M. Generating high quality crowd density maps using contextual pyramid CNNs [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1861-1870.